Final Project Report for IAL620-02: Text Mining & Natural Language Processing

Water Reports & News Classification using BERT Classification and sentiment analysis, Text mining, word embedding.

Pradnya Patil
Master's in informatics and Analytics
University of North Carolina, Greensboro, USA
papatil@uncg.edu

Problem Statement:

Now a days due to global warming and high amount of pollution, we are getting new information about environment and that is accessible to public on every second, the information such as reports, news, books, articles, etc. Using the water report and news, we can classify the data into different category and make the decision more efficiently.

In this project, we seek to answer the following data research questions:

A. Exploratory Data Analysis on the raw downloaded dataset along with text cleaning and transformation.

B. Explain different methods to analyze text and extract features that can be used to build a classification model

C. Classify water news and Reports based on subject text using Machine Learning Models (Decision Trees, SVM, Random Forest...).

D. Predict the probability of the different model of classification.

E. Do sentiment analysis of news and reports related to the water problems based on environmental issue.

Project Abstract:

In last years, water quality has been threatened by various pollutants and some environmental issues. Therefore, modeling and predicting water reports and news have become very important in controlling water pollution. In this project, models are built to classify the reports and news based on the 5 different target variables using the machine learning algorithms.

For this project, I am collecting the from https://www.kaggle.com/vbmokin/nlp-reports-news-classification. From this analysis, I will learn the key concept of the NLP mostly TEXT classification where the problem of assigning the news and report into their categories. Along with the Classification, I will learn the sentiment analysis also.

From this process, we can tell the reports, or the news are related to environment problems, pollution, treatment, climate indicators or the biotic monitoring in water or in a river basin.

Data Source and description:

The dataset contains 2 files so far - an English-language dataset from the English-language edition of the book, where the co-author, and a Ukrainian-language dataset from a separate Ukrainian-language edition of this book. These datasets contain approximately 95% of the same information.

Text Feature:

- Text - One or more sentences from reports or news

Binary Target Features:

- Env_problems - Is the text about an environmental problem? (0 or 1)
- Pollution - Is the text about environmental pollution? (0 or 1)
- Treatment - Is the text about treatment plants or environmental technologies? (0 or 1)
- Climate - Is the text about climatic indicators? (0 or 1)
- Biomonitoring - Is the text about biological, biotic monitoring in water or in a river basin? (0 or 1)

In this project I have used the following languages, frameworks, tools, libraries, packages for data extraction to Sentiment analysis and water reports & news data classification/regression.

Technologies: *R and Python*

Libraries:

R - tidytext, tidyverse, tm, lubridate, ldatuning, fliptime, stringr, ggmap, maps, stringr, rvest, quantenda, readtxt, textplots, widyr, ggplot2, igraph, ggraph

Python - numpy, pandas, matplotlib, seaborn, sklearn, torch, transformers, summarizer

Tools: Rstudio, Jupyter hub for python

Methods: R Selenium, BERT Classification

Water Report & News Data Description Table -

Attribute	Data Type	Description
text	String	One or more sentences from reports or news
env_problems	Integer	Is the text about an environmental problem? (0 or 1)
pollution	Integer	Is the text about environmental pollution? (0 or 1)
treatment	Integer	Is the text about treatment plants or environmental technologies? (0 or 1)
climate	Integer	Is the text about climatic indicators? (0 or 1)
biomonitoring	Integer	Is the text about biological, biotic monitoring in water or in a river basin? (0 or 1)

WSJ News Data Description Table -

Attribute	Data Type	Description
URL	String	The URL path of the news site
Title	String	The Title of the news from that URL
Date	String	The date of the news that published on what day, and by whom.
Text	String	All the Text contains the news that published by WSJ.

Data Analysis & Results –

News, reports, and articles are the most important unstructured data sources now a days along with the structured dataset. In the news, reports people describe about the health, food, world, environment, travel, and every topic. From this unstructured data form, we can in out the different result and do the data analysis and help to the world in different aspects.

Here are some of my analyses about the news and reports from my Exploratory data analysis.

Water Report and News EDA -

This is the graph that shows the missing values of your data set. In my Water Report and News data set I don’t have any missing values.

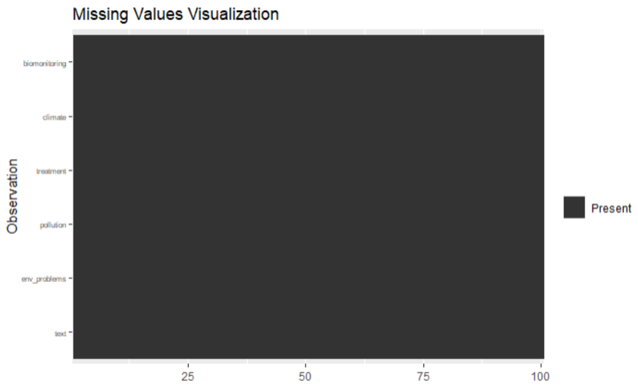
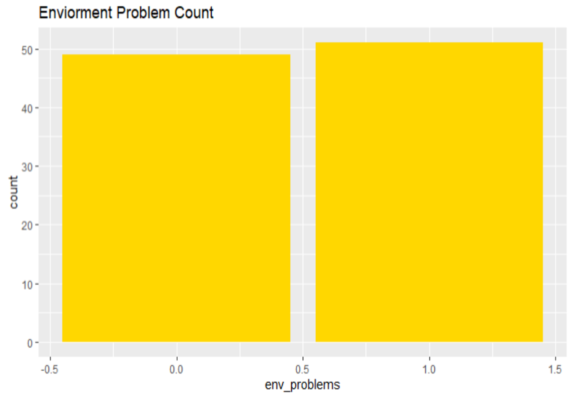
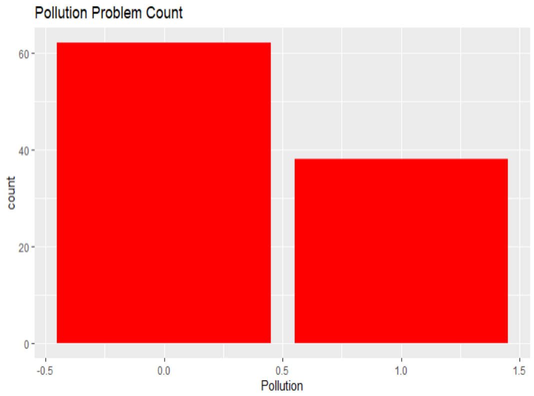
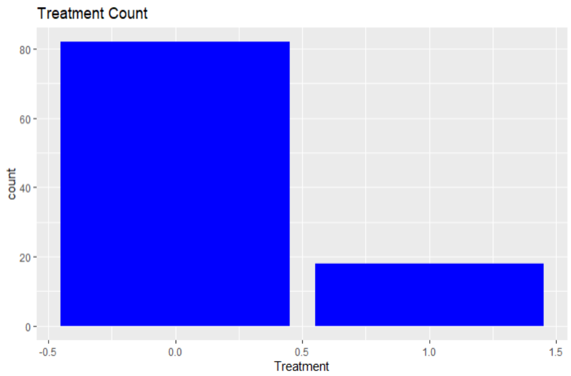


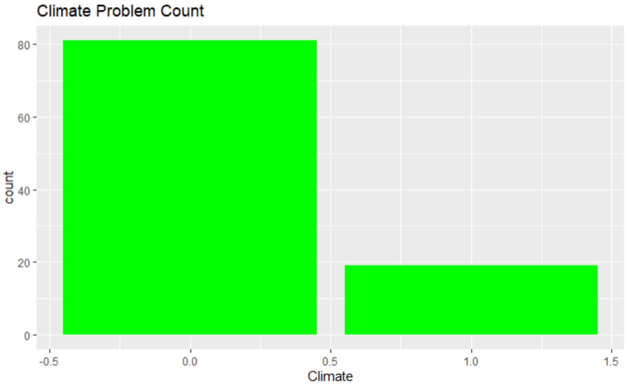
Figure 1

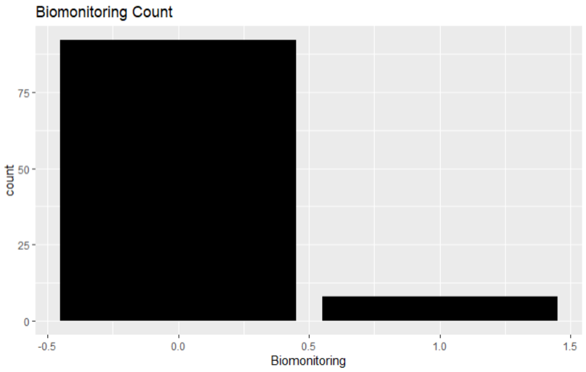
The below graphs are the bar plot of the target variables from the dataset.



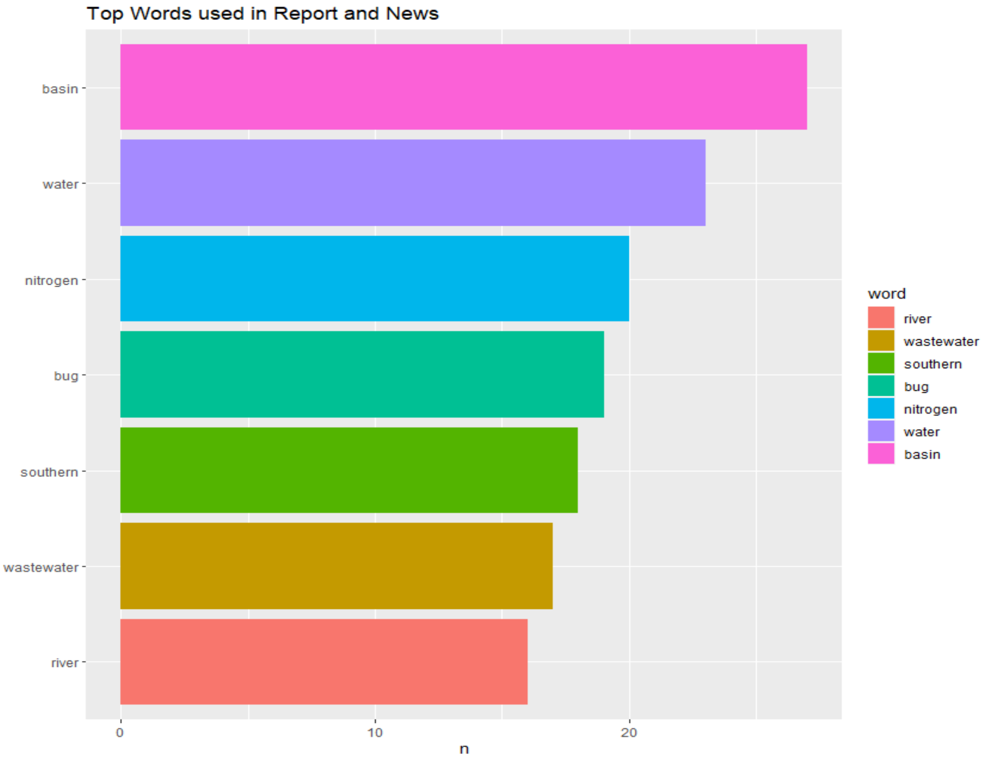


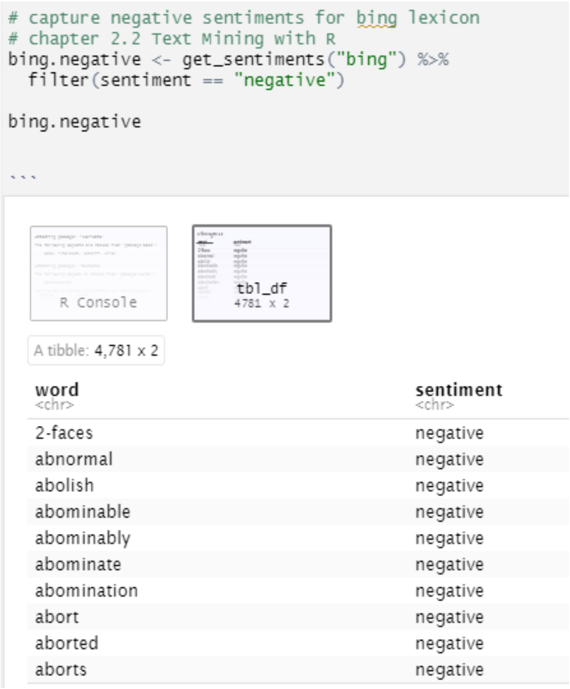
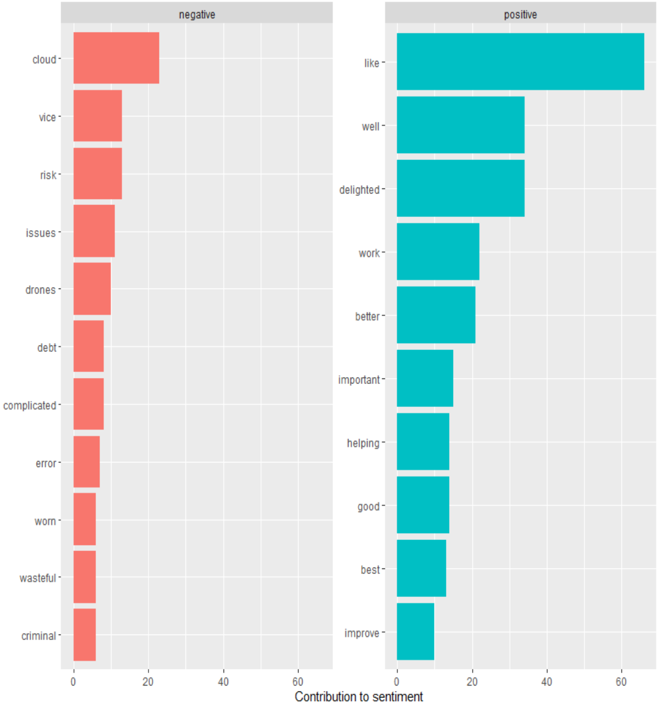


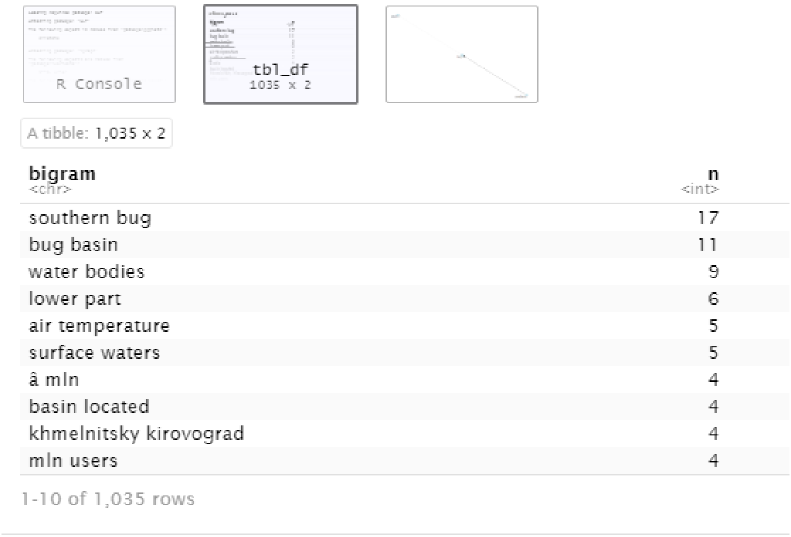




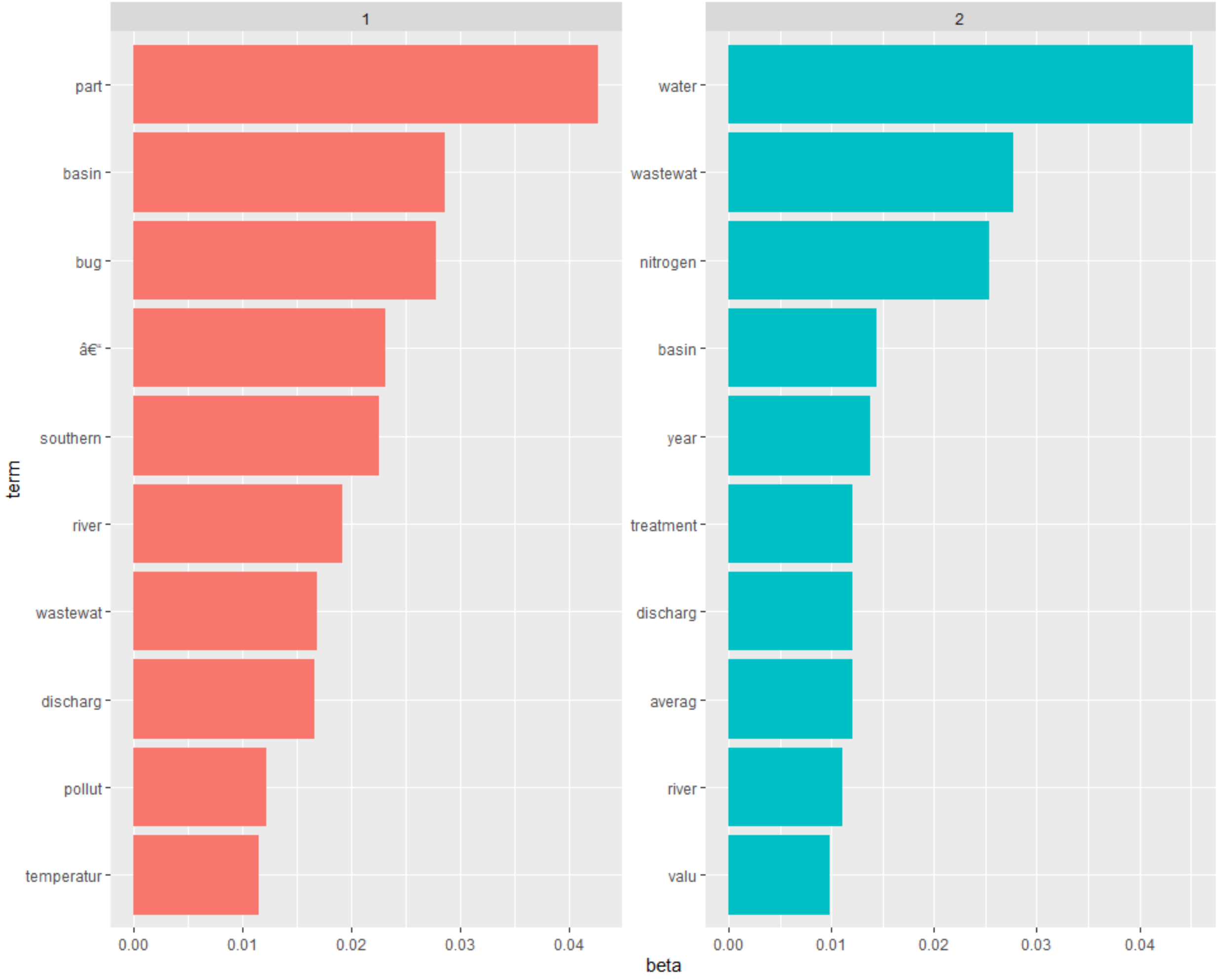
This is the diagram of the top words that have been used in the text of the dataset.

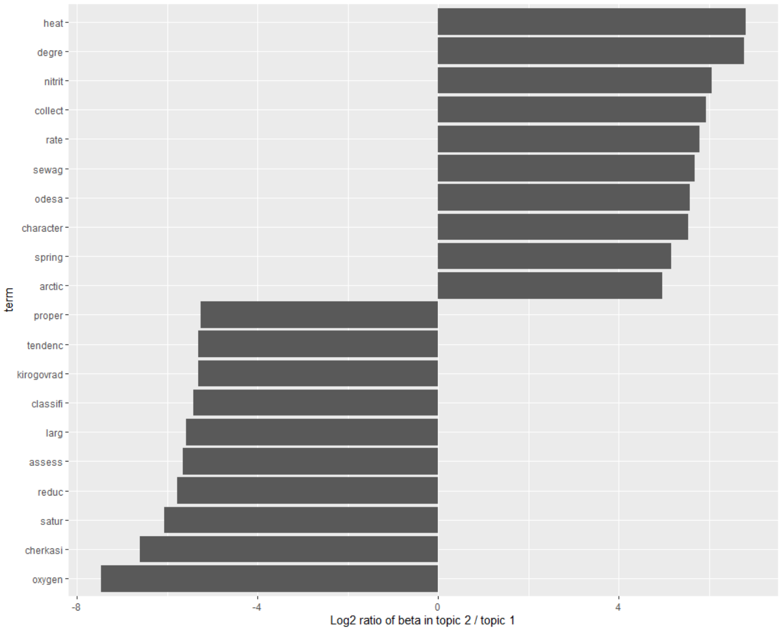



```
# capture negative sentiments for bing lexicon
# chapter 2.2 Text Mining with R
bing.negative <- get_sentiments("bing") %>%
  filter(sentiment == "negative")

bing.negative

```
```

A tibble: 4,781 x 2

word <chr>	sentiment <chr>
2-faces	negative
abnormal	negative
abolish	negative
abominable	negative
abominably	negative
abominate	negative
abomination	negative
abort	negative
aborted	negative
aborts	negative


```

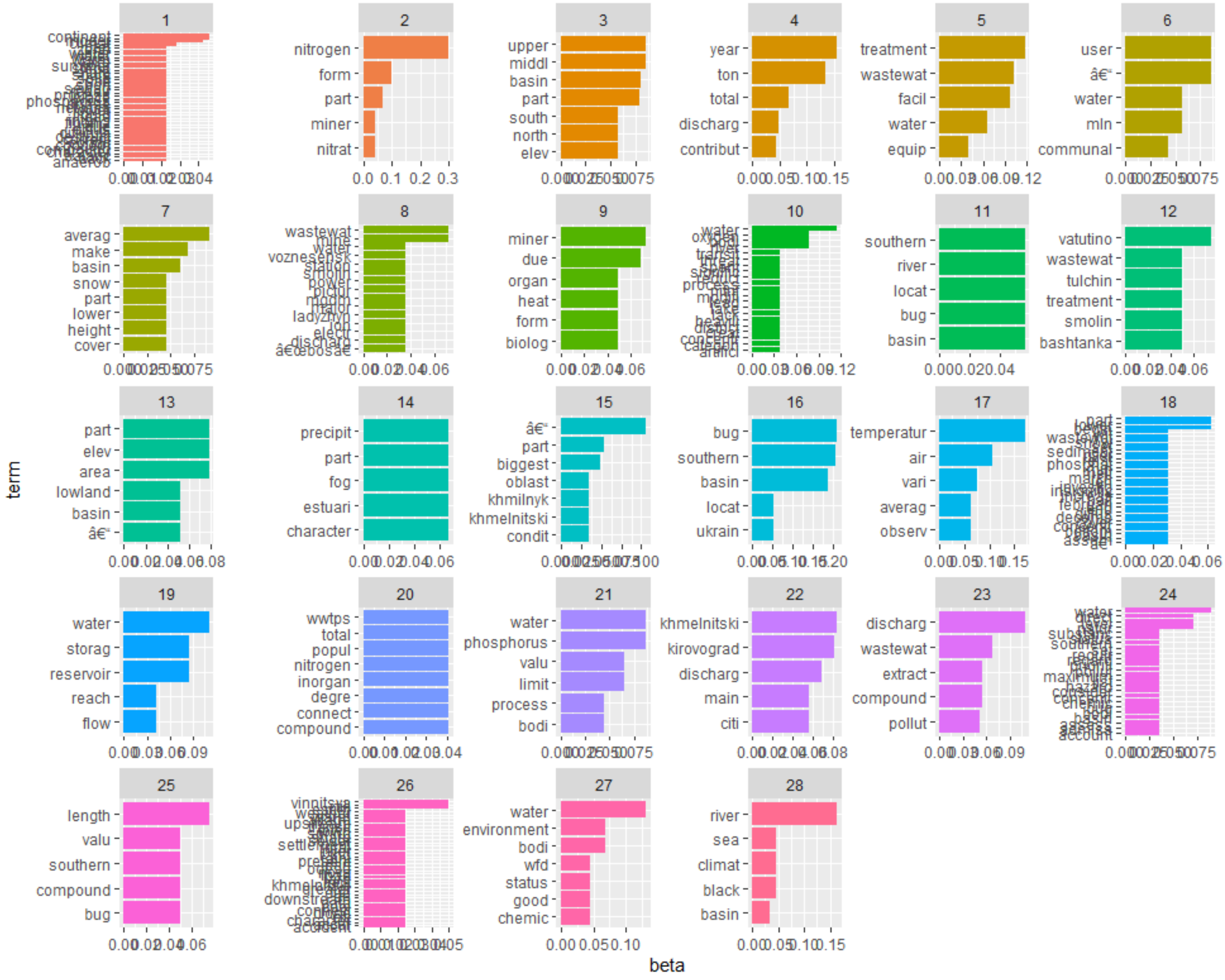


```
A tibble: 1,035 x 2
bigram                                   n
<chr>                                 <int>
southern bug                             17
bug basin                                11
water bodies                              9
lower part                                6
air temperature                           5
surface waters                            5
â mln                                     4
basin located                             4
khmelnitsky kirovograd                    4
mln users                                 4
1-10 of 1,035 rows
```



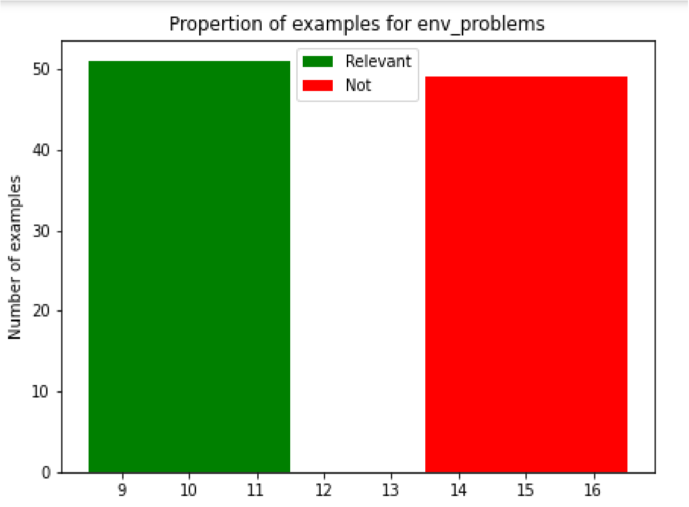


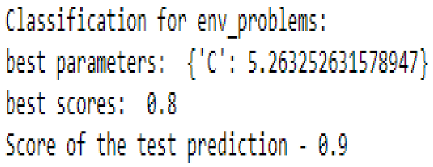
heat
degre
nitrit
collect
rate
sewag
odesa
character
spring
arctic
proper
tendenc
kirogovrad
classifi
larg
assess
reduc
satur
cherkasi
oxygen
term
-8
-4
0
4
Log2 ratio of beta in topic 2 / topic 1



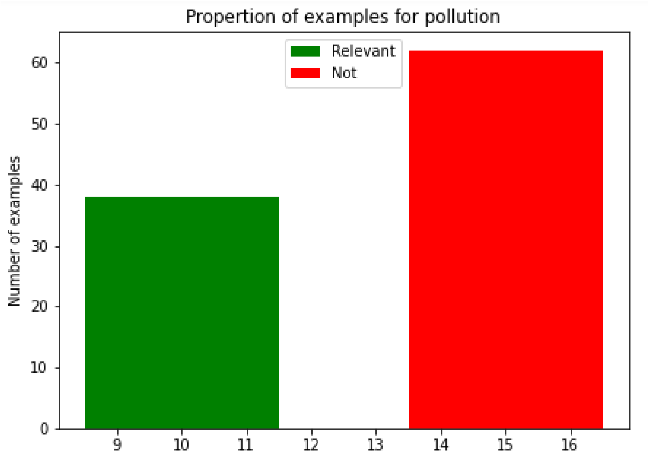
1
continent
phosphorus
anaerob
0.000.010.020.030.04
2
nitrogen
form
part
miner
nitrat
0.0 0.1 0.2 0.3
3
upper
middl
basin
part
south
north
elev
0.0000.0250.0500.075
4
year
ton
total
discharg
contribut
0.000.050.100.15
5
treatment
wastewat
facil
water
equip
0.000.030.060.090.12
6
user
â€"
water
mln
communal
0.0000.0250.0500.075
7
averag
make
basin
snow
part
lower
height
cover
0.0000.0250.0500.075
8
wastewat
mine
water
voznesensk
ladyzhyn
elect
discharg
â€œbosâ€
0.000.020.040.06
9
miner
due
organ
heat
form
biolog
0.000.020.040.06
10
water
oxygen
river
process
concentr
0.000.030.060.090.12
11
southern
river
locat
bug
basin
0.000.020.04
12
vatutino
wastewat
tulchin
treatment
smolin
bashtanka
0.000.020.040.06
13
part
elev
area
lowland
basin
â€"
0.000.020.040.060.08
14
precipit
part
fog
estuari
character
0.000.020.040.06
15
â€"
part
biggest
oblast
khmilnyk
khmelnitski
condit
0.0000.0250.0500.0750.100
16
bug
southern
basin
locat
ukrain
0.000.050.100.150.20
17
temperatur
air
vari
averag
observ
0.000.050.100.15
18
part
wastewat
sediment
phosphorus
0.000.020.040.06
19
water
storag
reservoir
reach
flow
0.000.030.060.09
20
wwtps
total
popul
nitrogen
inorgan
degre
connect
compound
0.000.010.020.030.04
21
water
phosphorus
valu
limit
process
bodi
0.0000.0250.0500.075
22
khmelnitski
kirovograd
discharg
main
citi
0.000.020.040.060.08
23
discharg
wastewat
extract
compound
pollut
0.000.030.060.09
24
water
southern
maximum
account
0.0000.0250.0500.075
25
length
valu
southern
compound
bug
0.000.020.040.06
26
vinnitsya
upstream
settlement
khmelnitski
downstream
account
0.000.010.020.030.040.05
27
water
environment
bodi
wfd
status
good
chemic
0.00 0.050.10
28
river
sea
climat
black
basin
0.000.050.100.15
term
beta

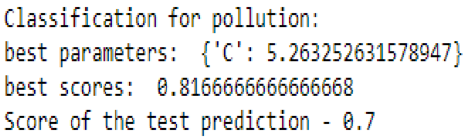
BERT Classification Results and Plots -



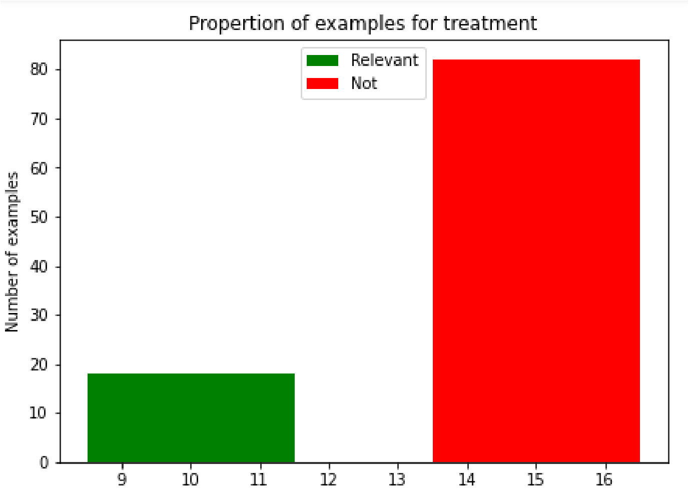


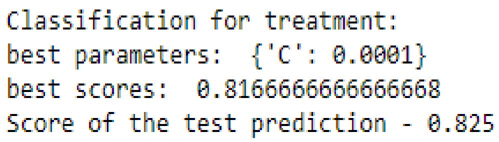
```
Classification for env_problems:
best parameters:  {'C': 5.263252631578947}
best scores:  0.8
Score of the test prediction - 0.9
```



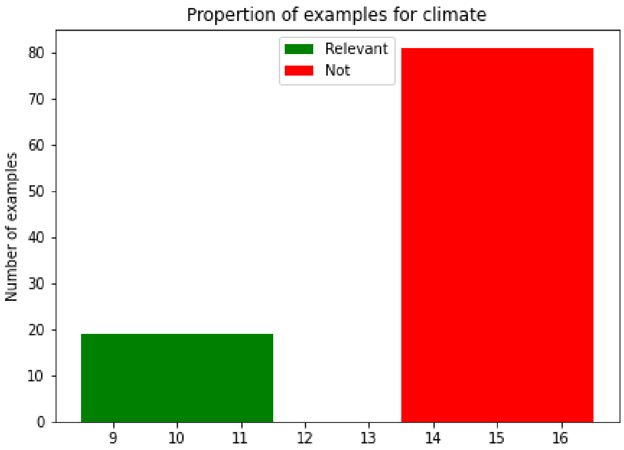


```
Classification for pollution:
best parameters:  {'C': 5.263252631578947}
best scores:  0.8166666666666668
Score of the test prediction - 0.7
```

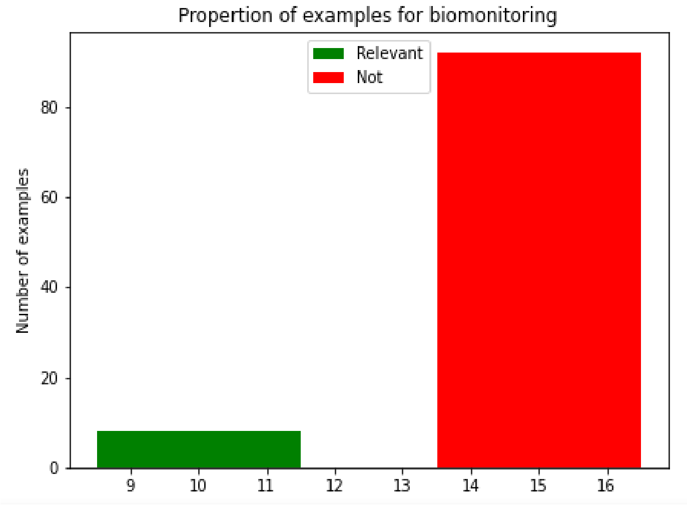




```
Classification for treatment:
best parameters:  {'C': 0.0001}
best scores:  0.8166666666666668
Score of the test prediction - 0.825
```

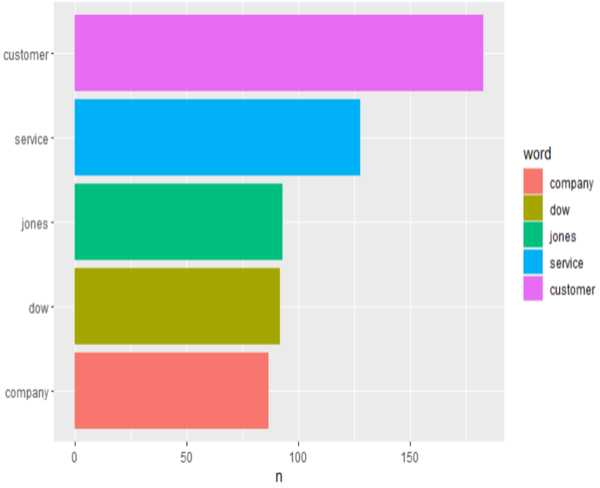


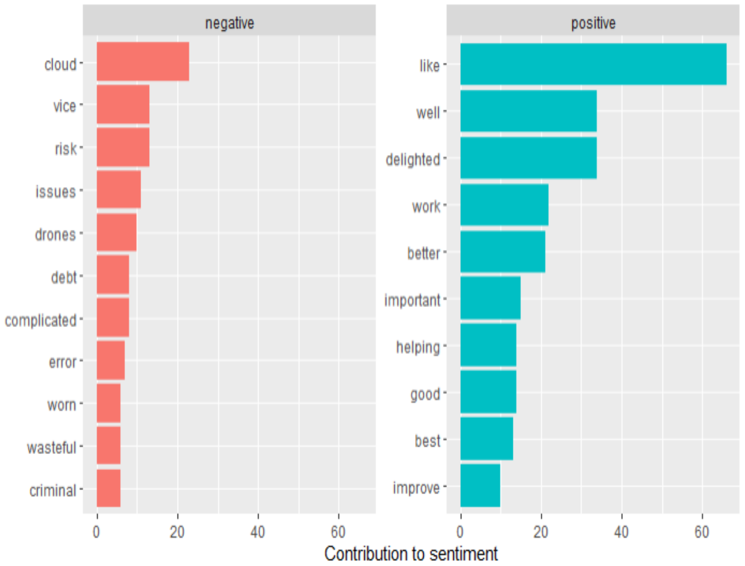
```
Classification for climate:
best parameters:  {'C': 5.263252631578947}
best scores:  0.9666666666666666
Score of the test prediction - 0.975
```

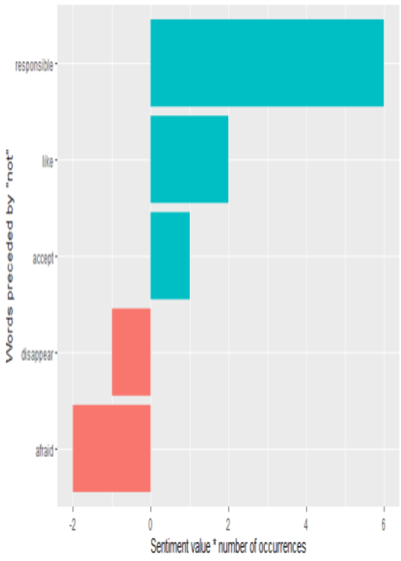


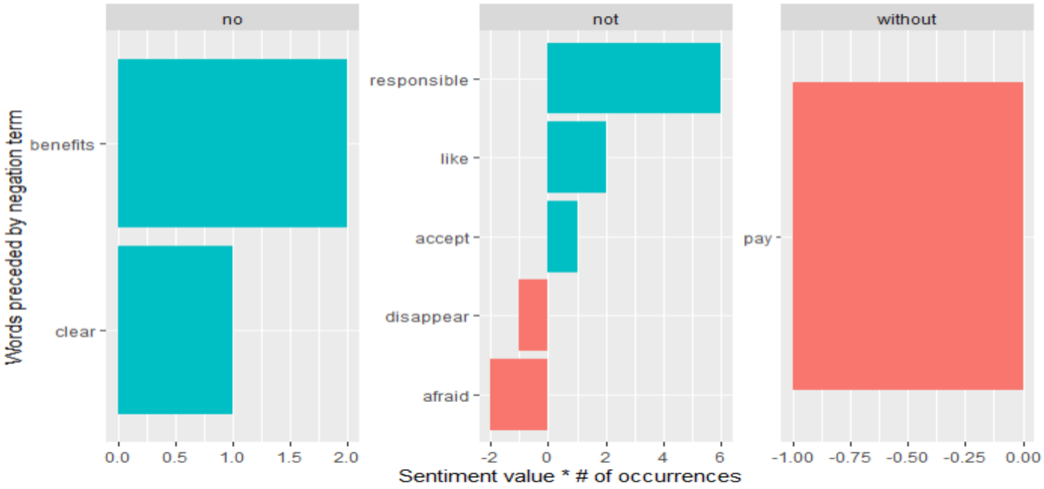
```
Classification for biomonitoring:
best parameters:  {'C': 0.0001}
best scores:  0.95
Score of the test prediction - 0.875
```

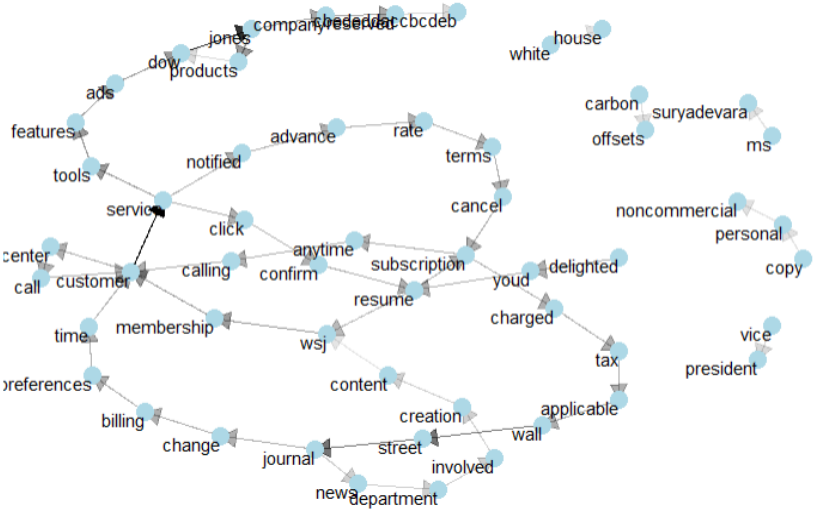
WSJ News EDA –

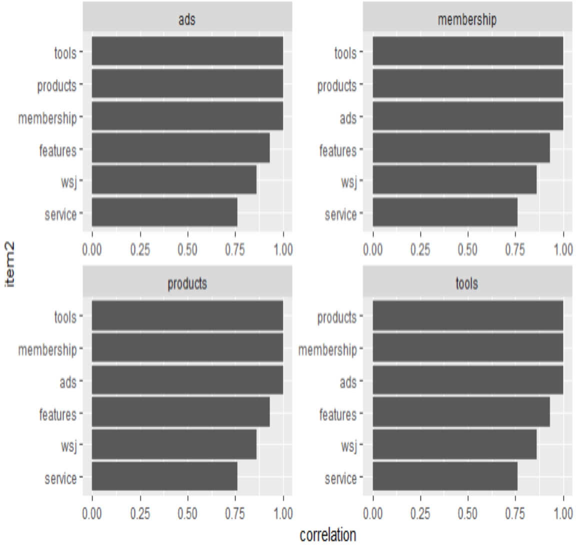


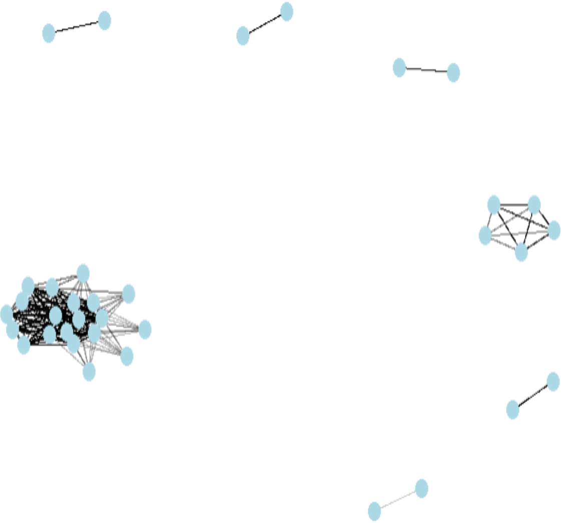

negative
positive
cloud
vice
risk
issues
drones
debt
complicated
error
worn
wasteful
criminal
like
well
delighted
work
better
important
helping
good
best
improve
0
20
40
60
0
20
40
60
Contribution to sentiment

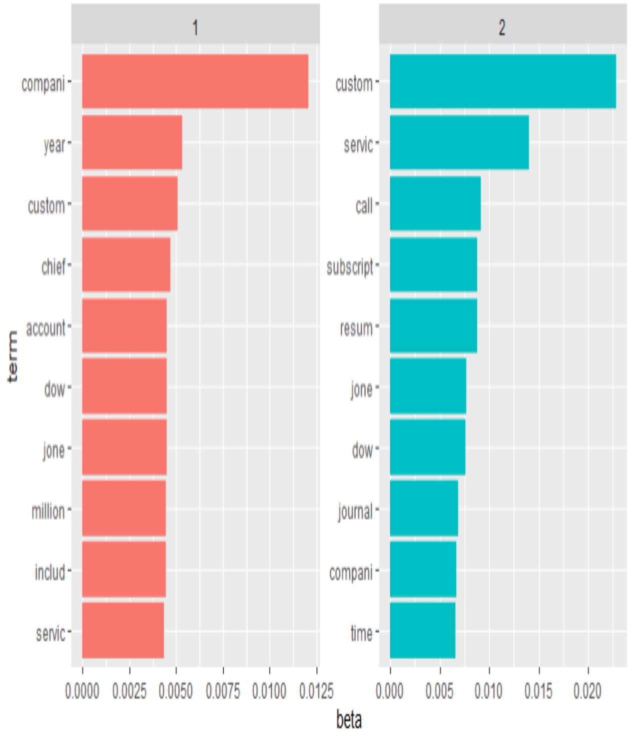

Words preceded by "not"
responsible
like
accept
disappear
afraid
-2
0
2
4
6
Sentiment value * number of occurrences

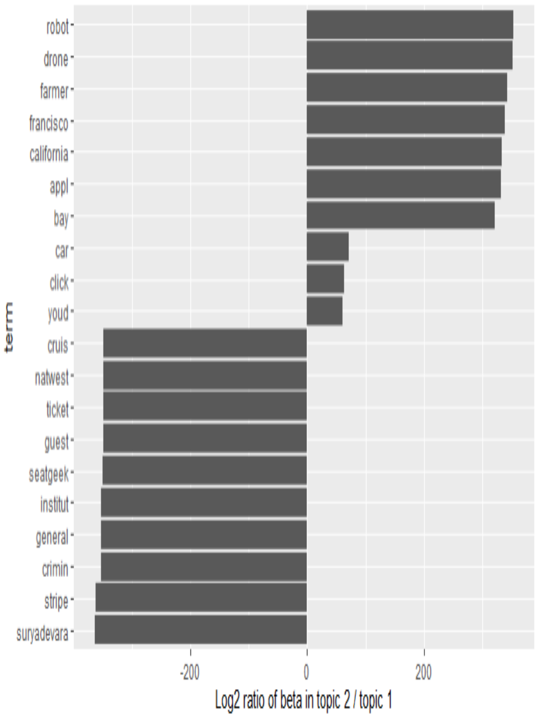

no
not
without
Words preceded by negation term
benefits
clear
responsible
like
accept
disappear
afraid
pay
0.0
0.5
1.0
1.5
2.0
-2
0
2
4
6
-1.00
-0.75
-0.50
-0.25
0.00
Sentiment value * # of occurrences

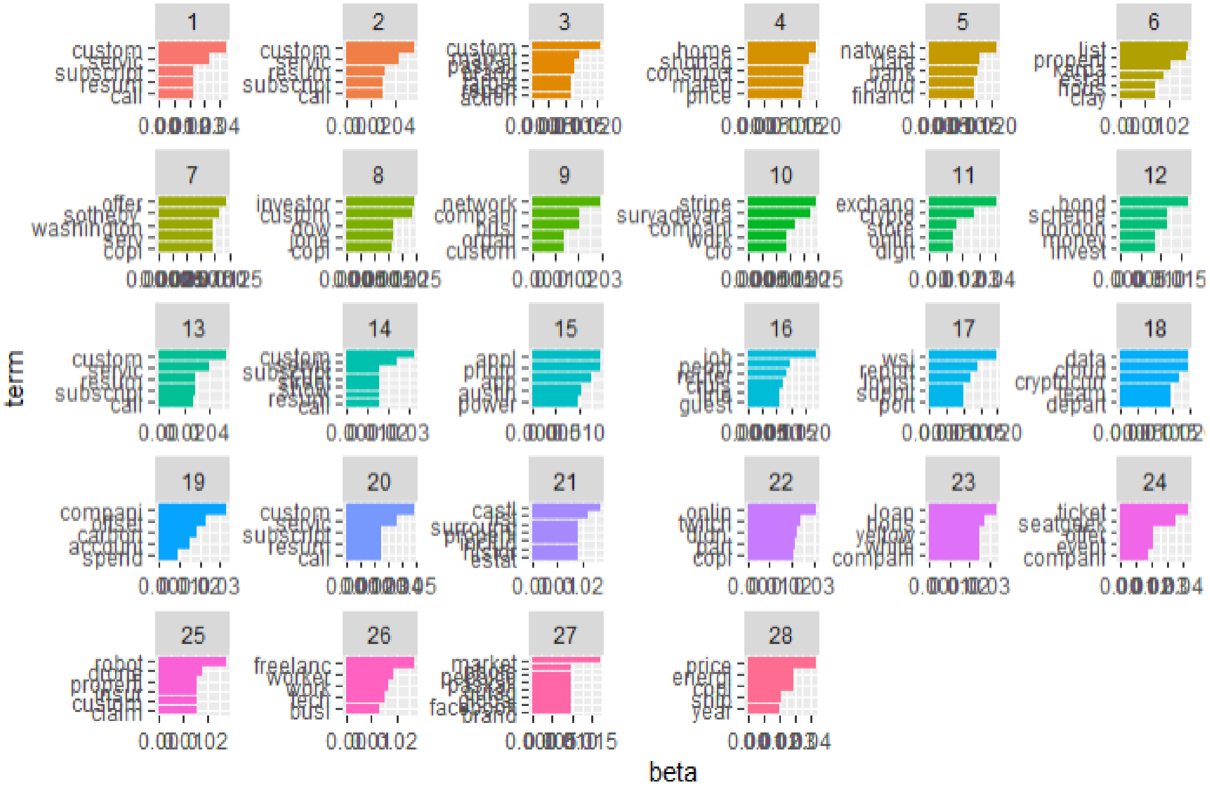

company
reserved
cbededdac
cbcdeb
jones
dow
products
ads
white
house
carbon
offsets
suryadevara
ms
features
advance
rate
terms
notified
tools
cancel
servic
click
noncommercial
personal
copy
center
call
customer
calling
anytime
confirm
subscription
youd
delighted
resume
charged
time
membership
wsj
tax
vice
president
preferences
content
billing
creation
applicable
change
journal
street
wall
involved
news
department

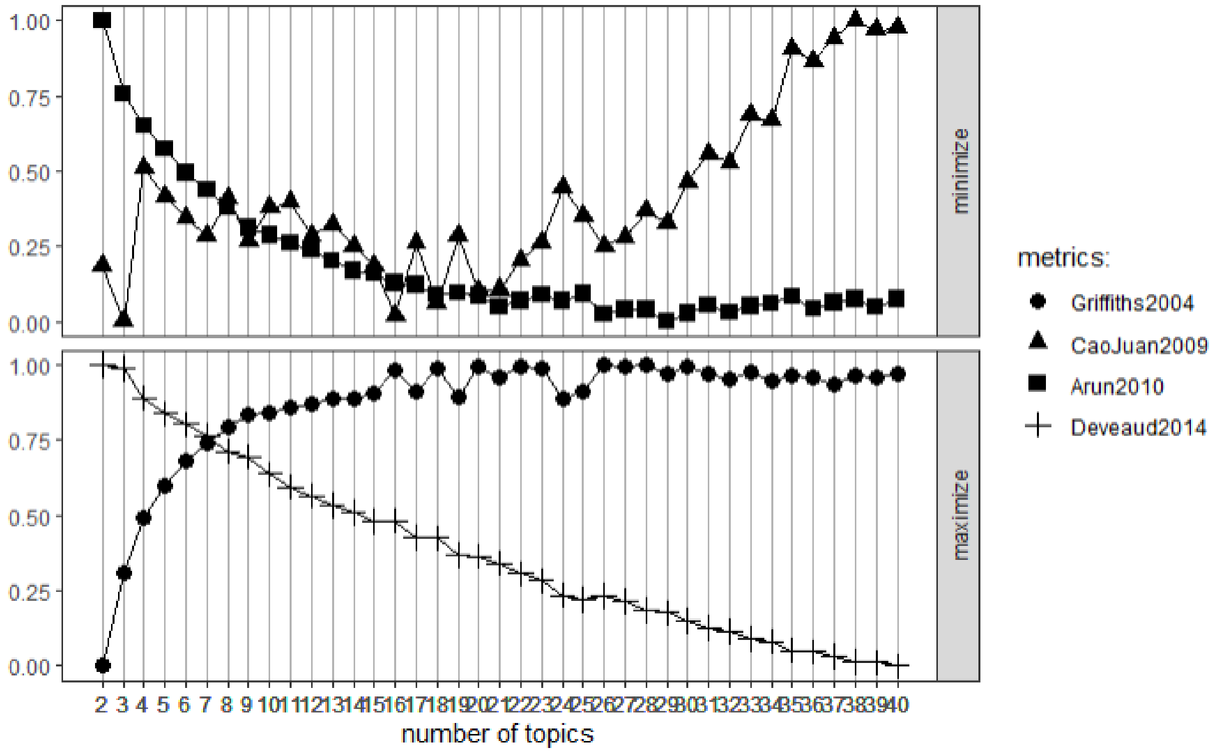

ads
membership
products
tools
tools
products
membership
features
wsj
service
tools
products
ads
features
wsj
service
tools
membership
ads
features
wsj
service
products
membership
ads
features
wsj
service
0.00
0.25
0.50
0.75
1.00
item2
correlation




1
2
compani
year
custom
chief
account
dow
jone
million
includ
servic
custom
servic
call
subscript
resum
jone
dow
journal
compani
time
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.000
0.005
0.010
0.015
0.020
term
beta


robot
drone
farmer
francisco
california
appl
bay
car
click
youd
cruis
natwest
ticket
guest
seatgeek
institut
general
crimin
stripe
suryadevara
-200
0
200
term
Log2 ratio of beta in topic 2 / topic 1


1
custom
servic
subscript
resum
call
2
custom
servic
resum
subscript
call
3
custom
action
4
home
shortag
construct
market
price
5
natwest
data
bank
cloud
financi
6
list
properti
kenya
estat
hous
clay
7
offer
sotheby
washington
copi
8
investor
custom
dow
jone
copi
9
network
compani
busi
oman
custom
10
stripe
suryadevara
compani
work
cfo
11
exchang
crypto
store
digit
12
bond
scheme
london
money
invest
13
custom
servic
resum
subscript
call
14
custom
subscript
resum
call
15
appl
phone
app
austin
power
16
job
guest
17
wsj
report
jobsi
support
port
18
data
cloud
cryptocurr
team
depart
19
compani
offset
carbon
account
spend
20
custom
servic
subscript
resum
call
21
cash
surround
properti
estat
22
onlin
twitch
copi
23
loan
yellow
compani
24
ticket
seatgeek
offer
event
compani
25
robot
drone
properti
custom
claim
26
freelanc
worker
work
tech
busi
27
market
facebook
brand
28
price
energi
coal
ship
year
term
beta


1.00
0.75
0.50
0.25
0.00
minimize
1.00
0.75
0.50
0.25
0.00
maximize
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
number of topics
metrics:
● Griffiths2004
▲ CaoJuan2009
■ Arun2010
+ Deveaud2014

Conclusion:

Github URL for code: https://github.com/Pradapatil/IAL62_NLP_-_TM

From this subject and Project, I have learned Text Mining and the Natural language processing methods. I did some data wrangling and scrubbing and find out the most frequent words that has been used in the reports and news.

From the Sentiment analysis learned how negative and positive words affected on our analysis. Using the word frequencies, N-gram an word correlation understood the concept and the relation between the words that has been used in text format.

Using the BERT classification model got the best score (0.96) for the climate variable and the got the highest prediction for the same variable which is (0.975)

References:

- Sklearn, Python library

References:

- https://www.kaggle.com/vbmokin/nlp-reports-news-classification
- https://towardsdatascience.com/text-analysis-feature-engineering-with-nlp-502d6ea9225d